Pipelines de ETL e Machine Learning com Apache Spark

Compreendendo o Conceito de SMOTE

O SMOTE (Synthetic Minority Oversampling Technique) é uma técnica amplamente utilizada para lidar com o problema de classes desbalanceadas em conjuntos de dados de classificação. Ele funciona criando novos exemplos sintéticos para a classe minoritária, aumentando assim sua representatividade sem simplesmente replicar os exemplos existentes. Vejamos como o SMOTE funciona.

Seleção de Vizinhos Próximos: Para cada exemplo da classe minoritária, o SMOTE identifica seus vizinhos mais próximos no espaço das características. Geralmente, isso é feito usando uma métrica de distância, como a distância Euclidiana, para identificar os k vizinhos mais próximos.

Geração de Exemplos Sintéticos: Novos exemplos são criados interpolando os valores das características entre o exemplo original e um de seus vizinhos próximos selecionados aleatoriamente. A interpolação funciona da seguinte forma: Para cada característica, o valor do exemplo sintético é gerado como uma combinação linear entre o exemplo original e o vizinho selecionado. Isso significa que o novo ponto estará em algum lugar ao longo da linha que conecta esses dois pontos no espaço de características.

Repetição do Processo: O processo é repetido até que a classe minoritária atinja o número desejado de exemplos, equilibrando o conjunto de dados.

Diferente de Oversampling Tradicional: O SMOTE evita simplesmente replicar exemplos existentes, o que pode levar ao overfitting. Em vez disso, ele introduz variabilidade ao criar novos exemplos no espaço das características.

Por ser baseado na interpolação, o SMOTE funciona principalmente com características numéricas. Em conjuntos de dados com variáveis categóricas, é necessário um pré-processamento adicional (como codificação) ou o uso de variações específicas do SMOTE. Existem extensões do SMOTE para lidar com diferentes cenários, como:

- Borderline-SMOTE: Gera exemplos sintéticos perto das fronteiras entre classes, onde o modelo pode ter mais dificuldade.
- ADASYN (Adaptive Synthetic Sampling): Foca mais nos exemplos minoritários que estão em regiões mais difíceis de classificar.
- SMOTE-NC: Funciona com dados mistos (numéricos e categóricos).

Recomendamos este material abaixo caso queira compreender a formulação matemática do SMOTE:

https://imbalanced-learn.org/stable/over_sampling.html#smote-adasyn

Equipe DSA

Muito Obrigado!
Continue Trilhando Uma Excelente Jornada de Aprendizagem.